Exploitation of Artificial Intelligence Based False Feedback Systems in Conjunction with Utility Data to Determine System Start Time and Therefore Position in Boot Process

13 January 2025
Simon Edwards
Research Acceleration Initiative

Introduction

For several years, artificial intelligence has been used in a defensive cybersecurity context not only in order to automate threat detection and mitigation, but to deceive intruders into believing that an intrusion was successful by providing the intruder with what appears to be an authentic root directory of a secure computer system. If an intruder falls for the deception, a counter-intrusion may be organized rapidly, typically using an automated process.

Abstract

The use of such systems, although they greatly enhance overall security, introduces a number of potential vulnerabilities, one of which is that these artificial-intelligence based security systems provide a source of constant feedback, even if that feedback is false. An intruder can, at minimum, deduce that the adversary network is powered on from the fact that there is any feedback. As this feedback is provided on a “must provide” basis, pings may be sent continually to determine whether a system is active. Although such computer systems are virtually always up and running, individual systems are brought offline for maintenance at semi-predictable times.

As computer systems; even in secure environments; universally feature the ability to remotely turn on or turn off the system, this means that the systems are necessarily listening for specific instructions such as a “Power On” instruction when the system is in the off position, so long as electrical energy is available to the system.

When one of the aforementioned false-environment generating firewalls (generally a separate piece of hardware,) are taken down for maintenance, this could be predicted to result in the system ceasing to provide even false feedback to the intruder. From the timing of the cessation of this false feedback, the intruder may deduce at what point the system is in its rebooting process, enabling genuine exploit code tailored to the firewall system’s most basic firmware to be sent to that system.

The inauthentic system/firewall may, in this way, be compromised and transformed into a physical source of a subsequent intrusion against the genuine system located behind the firewall.

As that sophisticated firewall comes back online, low-level firmware may be modified by an entity with knowledge of the specific point in the boot process at which the firewall currently exists. These vulnerabilities may be termed hyper-transient vulnerabilities and require specific knowledge of *when* a

system is being brought back online, accurate to within a margin measured in milliseconds.

Once within the firewall system, a multi-step process could allow the genuine system to be compromised, particularly in the context of multiple maintenance reboot cycles of the firewall system with a vulnerability being introduced each time a system is rebooted. With each reboot cycle, the level of control over the firewall system becomes greater. What begins as control only over power control firmware turns into control over the data bus and eventually becomes full system control in approximately three cycles.

The time at which this feedback ceases may not provide sufficiently accurate information concerning boot process position, however, when this information is used in conjunction with hyper-accurate information derived from a utility provider concerning spikes in power consumption, this combined information set can be leveraged to extrapolate the needed information. As the electrical utility provider's cybersecurity is likely to be trivial in comparison to the targeted system, this sort of data should be accessible.

Conclusion

General knowledge of planned system maintenance times, the more specific knowledge of false feedback cessation time and the hyper-specific timing information provided by the utility company describing an increase in power consumption attributable to the act of powering 'on' a firewall system may provide sufficient information for a hyper-transient vulnerability to be exploited successfully, permitting a foothold in even the most sophisticated firewall system.